A plea for more interactions between psycholinguistics and natural language processing research

Marc Brysbaert* marc.brysbaert@ugent.be
Emmanuel Keuleers* emmanuel.keuleers@ugent.be
Paweł Mandera* pawel.mandera@ugent.be

*Department of Experimental Psychology, Ghent University, H. Dunantlaan 2, 9000 Gent, Belgium

Abstract

A new development in psycholinguistics is the use of regression analyses on tens of thousands of words, known as the megastudy approach. This development has led to the collection of processing times and subjective ratings (of age of acquisition, concreteness, valence, and arousal) for most of the existing words in English and Dutch. In addition, a crowdsourcing study in the Dutch language has resulted in information about how well 52,000 lemmas are known. This information is likely to be of interest to NLP researchers and computational linguists. At the same time, large-scale measures of word characteristics developed in the latter traditions are likely to be pivotal in bringing the megastudy approach to the next level. We describe a recent evolution in word recognition research, which we think is of interest to natural language processing (NLP) researchers. First, we explain the nature of the new approach and why it has come to supplement (or maybe even replace) traditional psycholinguistic research (Section 1). Then, we describe how this has led to the collection of new word characteristics (Sections 2 and 3), which are likely to be useful for NLP researchers as well (Section 4). We end by illustrating how the new approach depended on input from NLP and needs further input to bring it to full fruition.

1. The transition from small-scale factorial designs to megastudies in psycholinguistics

Word recognition research has recently shifted to large datasets (Balota et al. 2004, Balota et al. 2007, Balota et al. 2013). Traditionally, this research was focused on small-scale factorial designs in which one or two variables were investigated while other variables were matched as much as possible. Before we discuss the new megastudy approach we present an example of the traditional factorial approach, so that the reasoning behind the new move becomes clear.

1.1 The traditional factorial design

A typical example of the factorial design is an experiment in which the effects of word frequency and word age of acquisition (AoA) are investigated. Gerhand and Barry (1999), for instance, wanted to examine a claim by Morrison and Ellis (1995) that word frequency no longer influences word processing efficiency once AoA is taken into account. To do so, Gerhand and Barry ran a lexical decision task. This is a task in which participants are shown a random sequence of words and made-up nonwords, and have to decide as rapidly as possible whether the presented letter string is an existing word or not. There were four types of words in Gerhand and Barry's experiment: early acquired low-frequency words, late acquired low-frequency words, early acquired high-frequency words, and late acquired high-frequency words. For each type, 16 words were selected so that they differed as much as possible on the two variables of interest and were matched on a number of control variables. For word frequency, Gerhand and Barry used two sources: Kučera and Francis (1967, American English) and Hofland and Johansson (1982, British English). AoA estimates were based on norms collected by Gilhooly and Logie (1980), who asked participants to estimate at what age they first learned each word, using a 7-point scale (where a rating of 1 was given to words

acquired between the ages of 0 and 2 years, and a rating of 7 was given to words acquired at age 13 and older). The control variables were word concreteness, imageability and length (number of letters). Figure 1 gives a summary of the stimulus characteristics.

Table 1
Characteristics of the Stimuli Used

	KF Freq.		HJ Freq.		AoA				
Stimuli	*M*	Range	*M*	Range	*M*	Range	Con.	Imag.	Length
Early, high-frequency (e.g., *cousin*)	206.3	51–847	197.9	19–953	2.67	2.19–2.92	5.05	5.25	5.6
Early, low-frequency (e.g., *rattle*)	4.2	0–9	4.6	0–17	2.71	2.19–2.97	4.92	5.35	5.6
Late, high-frequency (e.g., *union*)	146.2	57–382	121.1	40–206	4.82	4.50–5.39	4.52	4.95	5.9
Late, low-frequency (e.g., *marvel*)	3.3	0–9	2.6	0–13	4.91	4.42–5.52	5.03	5.14	5.6

Figure 1: Stimulus materials as used in a typical factorial psycholinguistic experiment. The two variables manipulated are word frequency (frequency per million words, coming from two different sources) and age of acquisition (rating from 1 to 7). The three control variables are word concreteness, imageability and length. Each of the four experimental conditions contained 16 words. Source: Gerhand & Barry (1999, Figure 1).

In addition to the words, 64 nonwords were created, so that participants could decide whether a presented letter string formed an existing word or not. The nonwords were created by using real words of the same length as each of the stimulus words and altering one or more letters. All nonwords were pronounceable, and none were homophonic to real words. Examples of the nonwords used were: *elt, hish, condim, and fashmoone.* Finally, 20 practice stimuli were created along the same lines, consisting of 20 medium-frequency words (with counts between 10 and 50 per million) of medium AoA (with ratings between 3 and 4.5), so that the participants had some experience with the task before they started the real experiment.

As Figure 2 shows, Gerhand and Barry (1999) found an effect of AoA as well as frequency on lexical decision times, indicating that both variables influence word processing. In addition, they observed an interaction effect, such that the frequency effect was larger for late acquired words than for early acquired words, or that the AoA effect was larger for low-frequency words than for high-frequency words. These findings undermined Morrison and Ellis's (1995) claim that the word frequency effect was an AoA effect in disguise.

1.2 Limitations of factorial designs

Factorial designs have been popular in psycholinguistics because they allow researchers to have a very precise look at the effects of isolated variables, even if these effects are small relative to the overall variability in the data, as is often the case in word recognition research. Indeed, it can safely be stated that a factorial design is the only way to investigate the contribution of a theoretically important variable that is expected to have but a small effect on overall processing times.

At the same time, the limitations of factorial designs are becoming clear, as can easily be illustrated with Gerhand and Barry's (1999) study.

1. **Extreme words may be exceptional.** By looking exclusively at the extreme values of a dimension, researchers may be focusing too much on stimuli that are not representative for the entire continuum. For instance, the first acquired words all involve references to the world of a toddler, particularly if they are of low frequency in corpora. Examples of these words in Gerhand and Barry (1999) are: *fairy, fisherman, berry, rattle, peep, knitting, tablespoon, vase.* Similarly, many of late acquired high-frequency words are related to studying and sciences. Examples are: *student, union, science, president, degree, professor.* This is the more a problem

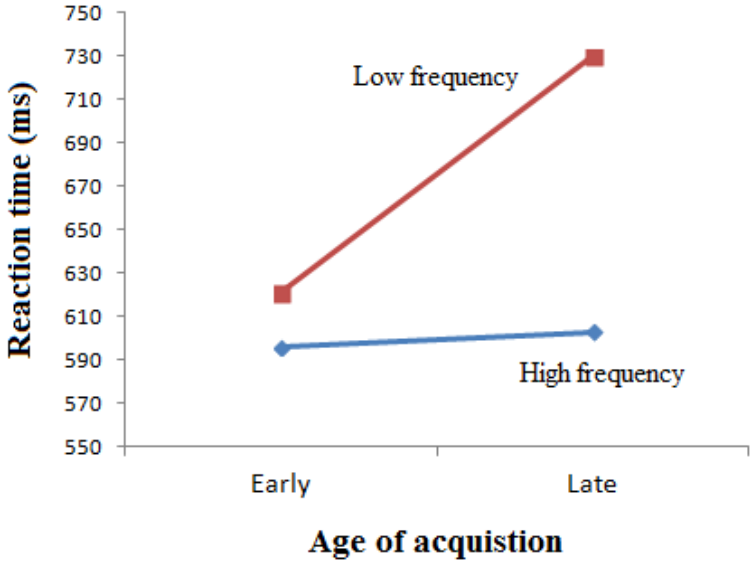

Figure 2: Results obtained by Gerhand & Barry (1999, Experiment 1), suggesting that both AoA and frequency affect lexical decision times and, in addition, interact with each other.

because the number of words per condition is rather small (only 16), putting a lot of weight on a few observations.

2. **No information about the influence of the variable across the entire range.** A factorial study gives very sparse information about the influence of the variable across the entire range. In general, it only provides two reference points between which a linear relationship is assumed (as shown in Figure 2). In addition, because there is no information about the curve of the distribution, the end points are selected on the basis of very little information. A look at the frequency values used by Gerhand and Barry (Figure 1) illustrates that they considered high-frequency words to have a frequency of more than 50 per million and low-frequency words to have a frequency of less than 10 per million (given that the Kučera and Francis frequencies were the norm in the 1990s, it is likely that stimulus selection was based on them and that the comparison with the Hofland and Johansson measures was added later).

3. **It is difficult to take into account all important control variables.** Gerhand and Barry's study illustrates another problem: How can we be sure that all relevant control variables have been taken into account? Apart from the fact that the words came from different realms of life, as illustrated above, there are two other variables not taken into account, which have been shown to influence word processing times (Brysbaert et al. 2011). These are the number of syllables in the word and the similarity to other words (with various measures, such as the number of orthographic or phonological neighbors, the Levenshtein distance with the closest neighbors, or the bigram frequencies of the letters). Apart from these two variables,

there is a plethora of other variables that at some point have been claimed to influence lexical decision times and that ideally should be controlled as well (Cutler 1981).

4. **No information about the relative importance of the variables.** One reason why there is no consensus about which variables to control is that factorial designs give very little information about the relative importance of the variables. All they provide, is whether the effect is statistically significant or not. The data of Figure 2 suggest that the effect of frequency is similar to that of AoA. However, all here depends on the end points that were chosen: To which extent are they comparable? As we will see below (Figure 3), Gerhand and Barry (1999) missed more than half of the frequency effect by only using words with frequencies higher than 1 per million words.

5. **A large proportion of unusual words are crammed together in a psycholinguistic experiment.** Another problem with factorial designs is that most of the time participants are confronted with many rare words in a short experiment. For instance, more than half of the words in the Gerhand and Barry study were words encountered at best a few times in a year. It is not clear to what extent this influences the decision criteria participants adopt to separate the words from the nonwords. In the worst case, this could lead to processing strategies no longer representative of normal word processing.[1]

6. **Word characteristics cannot be manipulated experimentally.** The main reason for the above problems is that word characteristics are stimulus specific. Ideally, in an experiment one can assign stimuli to one or the other condition at random or in a counterbalanced design. This is the case, for instance, in semantic priming experiments, where target words (e.g., *doctor, cat*) are preceded by related primes (*nurse, dog*) or unrelated primes (*purse, log*), and one investigates how much faster the target words are recognized when they are preceded by related primes (*nurse-doctor, dog-cat*) than when they are preceded by unrelated primes (*purse-doctor, log-cat*). Critically, in such experiments the experimenter has full control over which words are presented in the related and unrelated conditions. So, some participants will see the combinations *nurse-doctor* (related) and *log-cat* (unrelated), whereas other participants will get the sequences *purse-doctor* (unrelated) and *dog-cat* (related). In this way the semantic priming effect is not confounded by the target stimuli in the two conditions (across participants the same target words are presented in the related and the unrelated conditions). Such counterbalancing is not possible in studies investigating the properties of words themselves. Gerhand and Barry could not assign the words at random to the various conditions (let alone counterbalance them); all they could do was *select* the words in the various conditions. This implies that their study was not a real experiment but a correlational study in disguise. Any effect found between the conditions could be a function not only of the variables manipulated but also of possible confounds between the words in the various conditions not taken into account.[2]

1.3 The alternative: Regression analyses of large numbers of words

The alternative to factorial designs is to collect data for many (ideally, 'all') words and run regression analyses on them. This type of analysis has been promoted, among others, by Baayen (2010), who concluded that "For predictors that are part of a complex correlational structure, dichotomization

1. For what it is worth, on the basis of our experiences we have the impression that the effects of variables are larger in lexical decision experiments containing only a few words with extreme values of a variable than in experiments where the words are embedded in a larger variety of stimuli, arguably because participants tune into the word features that distinguish them from the nonwords.
2. The equivalent in semantic priming would be an experiment in which one compares words with related primes to other words with unrelated primes (e.g., the prime-target pair dog-cat in the related condition is compared to the prime-target pair log-bat in the unrelated condition; in such an experiment the difference between the related and the unrelated conditions is not only a function of the type of prime used, but also of the target words used).

almost always leads to a loss of statistical power. For such predictors, a 'real' experiment is not a factorial experiment but a regression experiment."

The regression approach with large numbers of words was initiated by Balota et al. (2004)(see also Spieler and Balota 1997) , who collected lexical decision times and word naming times for 2,428 monosyllabic English words, an enterprise that was later extended to 40,000 words by Balota et al. (2007). The collection of word processing data for a large number of unselected stimuli is known in psycholinguistics as the *megastudy approach*. The database for English compiled by Balota et al. (2007) is called the *English Lexicon Project*.

The availability of processing times for large numbers of words makes it possible to examine the influence of variables across the entire range of values. In addition, one is no longer limited to linear regression analysis. Keuleers et al. (2010a), for instance, mapped the word frequency effect using nonlinear regression (Figure 3). They observed that the word frequency effect was indeed more or less linear for frequencies between 1 per 100 hundred million words and 10 per million words, but leveled off for higher values. In addition, nearly half of the effect was situated below frequencies of 1 per million. The frequency effect was nearly absent for frequencies above 100 per million (in other studies, this could go as low as 50 per million). So, if Gerhand and Barry had included words with frequencies lower than 1 per million words in their study, they would have found a larger frequency effect than the one shown in Figure 2.

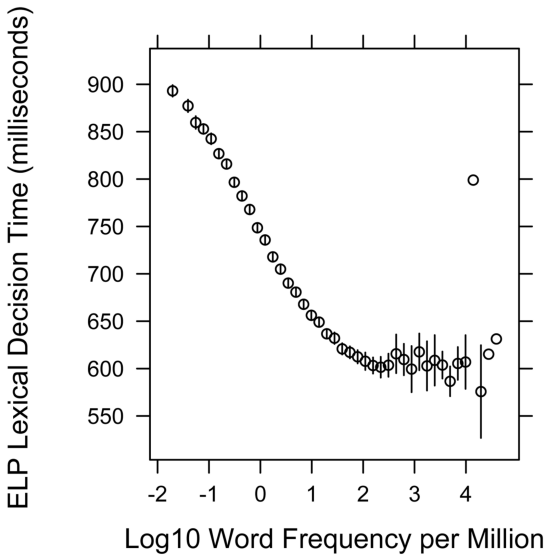

Figure 3: The word frequency effect in the English Lexicon Project lexical decision data. The curve shows that above a frequency of 100 per million (log10 = 2), there is no more difference between the stimuli. In contrast, nearly half of the frequency effect is situated below frequencies of 1 per million (log10 = 0). Source: Keuleers et al. (2010a).

Along the same lines, authors have looked at the effect of AoA across the entire range (Kuperman et al. 2012) or the effects of word valence and word arousal (whether a word is experienced as positive (*sunshine*) or negative (*molester*), and how much arousal it provokes (low: *grain*; vs. high: *lover*); Kuperman et al. 2014). Because of the usefulness of the data from the English Lexicon Project, similar megastudy data have been collected in Dutch (Keuleers et al. 2010a), French (Ferrand et al. 2010), British English (Keuleers et al. 2012), and Chinese (Sze et al. 2014). As the numbers of stimuli were considerable, this also required the availability of a good automatic pseudoword generator (Keuleers and Brysbaert 2010).

2. A need for more and better word norms

2.1 Better measures of word frequency

The availability of processing times for large numbers of words in turn increased the need for more and better word norms. One of the first uses of the megastudy data indeed was to test the usefulness of various word frequency measures (Balota et al. 2004, Brysbaert and New 2009). For the first time, lexical decision times for thousands of words could be used as a validation criterion to see how well the word frequencies predicted lexical decision times. It immediately became clear that the widely used Kučera and Francis (1967) measure was much less predictive than more recent measures, partly because of the small size of the corpus (only one million words, making it impossible to account for the frequency effect below 1 per million words). It also became clear that word frequencies based on film subtitles were better than word frequencies based on books and newspapers (Brysbaert and New 2009, Brysbaert et al. 2011, Cai and Brysbaert 2010, Dimitropoulou et al. 2009, Keuleers et al. 2010a, Mandera et al. in press b, New et al. 2007, van Heuven et al. 2014). This has led to the creation of so-called SUBTLEX word frequencies for various languages.

2.2 Subjective norms for large samples of words

Another shortage that became felt was the limited availability of subjective word norms, the most important being AoA measures, measures of concreteness, and measures of the affective strength of the concepts referred to by the words. Having subjective ratings for a few words only becomes frustrating when one has access to word processing times and frequency measures for tens of thousands of words. As a result, large-scale norming studies have been designed to collect these data. In American English, this was facilitated by the availability of Amazon Mechanical Turk, a service making it possible to contact thousands of users who are willing to provide word ratings for a feasible price. As a result, subjective norms for AoA, concreteness and affective values have become available for most of the English words (Brysbaert et al. 2014a, Kuperman et al. 2012, Warriner et al. 2013).

The situation is more complicated in other languages, as Amazon Mechanical Turk is limited to the US. However, recent research in Dutch (Brysbaert et al. 2014c, Moors et al. 2013) showed that valid ratings can be obtained by asking participants to provide ratings for up to 6,000 words, if they are paid reasonably well and given enough time and freedom to complete the list. The costs are the same, but the logistics become more feasible, as one can collect ratings for 30,000 words with a group of 100 participants (20 participants per list, 5 lists).

2.3 More systematic collection of word stimuli

As databases grew larger, there was an increased need for systematicity in the word lists used. Traditionally, psycholinguists relied on word lists based on corpus research (i.e., the words included in word frequency lists). However, for two reasons this was felt to be suboptimal. The first problem is that word lists based on corpus analysis include all types of non-interesting word types (inflected forms, transparent compounds and derived forms, proper nouns, typos, and so on), which increase the costs of data collection. Indeed, a corpus of a few hundred million words easily provides a list of more than 500,000 types, only some of which are interesting for a rating study. The second problem is that there is an element of circularity if stimulus lists are exclusively based on word frequency lists. In such lists, it can be expected that the contribution of word frequency will be overestimated, because words absent from the frequency list (i.e., with a frequency of 0) are left out of consideration.

Ideally, one would have access to the full list of words in a language, as provided by the most prestigious dictionary. Unfortunately, for commercial reasons publishers of dictionaries are unwilling to provide these data (or at least will seriously limit the use of them[3]). Another option is to compile

3. For instance, it is unlikely that the word list can be made available freely to other researchers, as this violates the publisher's copyrights.

a list oneself on the basis of different corpora and freely available word lists collected by others (e.g., catalogs of shops). These lists are likely to miss some interesting words known to a large proportion of the population, but can approach the ideal if they are updated each time a new source becomes available.[4] Thus far we made such attempts for Dutch and English.

3. Crowdsourcing to find out which words are known

3.1 The Groot Nationaal Onderzoek [Big National Research] initiative

Having a 'full' list of words in a language is not so interesting in psycholinguistics. More important is to know which words are familiar to people and likely to be used by a sufficiently large proportion of the population. These are the words that really matter. An opportunity to collect such information arose when we were contacted by the Dutch broadcasting companies NTR and VPRO, who wanted to run a nation-wide study (as part of their Groot Nationaal Onderzoek [Big National Research] program). We took inspiration from the yes/no vocabulary test developed for second language proficiency (Lemhöfer and Broersma 2012, Meara and Buxton 1987). In this test, very similar to a lexical decision task, a number of words are presented among nonwords (typically in a 2:1 ratio)[5] and participants have to indicate which words they know. A penalty is given for nonwords wrongly selected, so that participants are encouraged only to accept those words they are (reasonably) certain about.

In our version of the test, each participant received a random sample of 100 stimuli, roughly two thirds of which were words and one third nonwords (for more information, see Brysbaert et al. 2014b, Keuleers et al. in press). To make the test rewarding for the participants, feedback about their performance was given in the form of the percentage Dutch words they were estimated to know. This was calculated as the percentage of words selected minus the percentage of nonwords erroneously picked. As a result of the media publicity and the feedback we provided (which could easily be shared by participants on the social media), the test went viral and after 8 months was completed over 650,000 times. Participants could do the test several times (as different stimuli were selected each time), so that the total number of participants was smaller than 650K. Still it can be estimated we reached about 2.5% of the Dutch-speaking population.

3.2 A new variable: Word prevalence

Because so many participants took part in the crowdsourcing study, we had on average 800 observations per word. This allowed us to calculate the percentage of people who know each word, a variable we call word prevalence (Keuleers et al. in press). Figure 4 shows the correlation between word prevalence (percent known) and the SUBTLEX word frequencies, which have the highest correlation with the lexical decision times of the Dutch Lexicon Project (Keuleers et al. 2010b, Keuleers et al. 2010a).

As expected, there was a positive correlation between word frequency and word prevalence. However, the correlation was rather modest (r = .50, meaning that only 25% of the variance in word prevalence was predicted by log frequency). In particular, we noticed that of the 52,800 words we presented, 22,000 were not in the SUBTLEX corpus of 43.7 million words (see the black line at the left side of the graph). More importantly, of these 22K words about half were known to more than 75% of the participants. Many of these words were compounds or derived words, such as *akkerbouw, baanbreker, bestuiving, bouwgrond, deelwoord, flitspaal, globaliseren, gospelmuziek, ham-*

4. Since the initial compilation of the Dutch word list of 52,800 words in January 2013, we had to omit some 2,200 less interesting entries (mostly outdated compounds) and added more than 6,000 new entries we came across.
5. The 2:1 ratio is chosen because few participants know all words. Depending on their vocabulary size, the ratio rapidly drops to the typical 1:1 ratio and even lower for participants with limited vocabulary. When the share of nonwords is larger than that of words, people start reinterpreting the task in such a way that a no-answer becomes the default response (Keuleers et al. 2012).

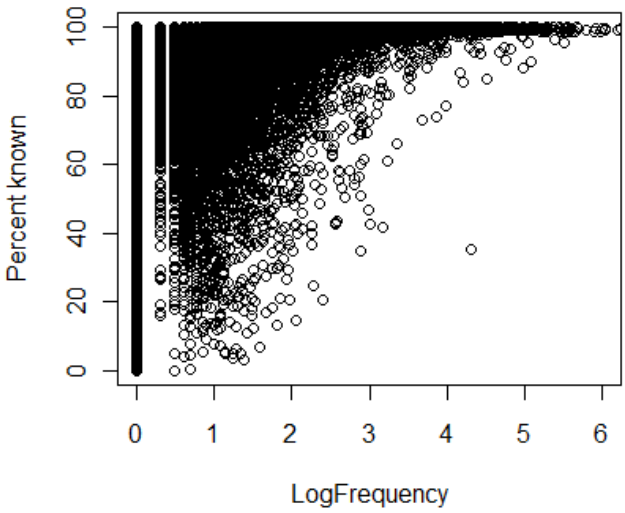

Figure 4: Correlation between word frequency (SUBTLEX) and word prevalence (percentage of words known in the crowdsourcing study Groot Nationaal Onderzoek) for the 50+K Dutch words included in the test. LogFrequency is based on 1 plus the number of observations in a corpus of 43.73 million words.

steraar, kijkcijferkanon, oppositiepartijen, overheidstaken, postpakket, proeflokaal, puntbaard, ramptoerist, rechtsbeginsel, regeerakkoord, scheurkalender. This shows the importance of not starting exclusively from the SUBTLEX list. Otherwise, these words would have been unavailable for research (and the word frequency effect on word prevalence would have been overestimated). Similarly, many of the words present only once or twice in the SUBTLEX corpus were known to nearly all participants. Examples include: *aanbidster, verfpot, zwerftocht, klapstoel, regenwolk, kaasplank, schietgebed, dorpsgenoot, trekvogel, graanproduct, bierglas, inleidend, smaakstof, kernwoord, kortharig*.

So, especially for the low-frequency words there are considerable deviations between word frequency and word prevalence (the upper left corner of Figure 4).

4. Use of the new data for NLP research

Needless to say, the turn towards megastudies and big data collection makes psycholinguistics more interesting for NLP researchers and computational linguists than the traditional, small-scale factorial designs. These are some of the uses we see (all our data are freely available at our website):

1. **Analysis of the subtitle corpora.** Word frequencies based on subtitles predict lexical decision times better, arguably because the language used in films is closer to everyday speech than the language used in written sources, in particular sources based on scientific and encyclopedic texts. It will be worthwhile to investigate whether the same is true for other NLP measures, such as word association values. A limiting factor with subtitle corpora is that film producers may have some copyright over the contents, so that the corpora cannot be made available for download on the internet (users only get access to them under the condition of fair use).

2. **The use of lexical decision data as a validation criterion.** Above we saw how the lexical decision times from the English Lexicon Project and the Dutch Lexicon Project were used to evaluate the quality of the existing word frequency measures. The same can be done for other word variables. For instance, estimates of orthographic and phonological similarity to other words can be based on different parameters and on different segments of the vocabulary (e.g., all words vs. all words known by more than 95% of the participants). It will be interesting to see whether there are considerable differences in the percentage of variance accounted for between the various estimates and whether there is some cross-language convergence on which measure is optimal.

3. **Word norms as golden standards.** Having human data to start from is ideal to test various algorithms meant to simulate the human data. For instance, our data on the affective values of words (Warriner et al. 2013) are heavily used in algorithms for text sentiment analysis (Guerini et al. 2013, Muresan et al. 2013, Ruppenhofer et al. 2014). Similar uses can be foreseen for the AoA ratings (Vajjala and Meurers in press) and the concreteness ratings (e.g., Hill and Korhonen 2014, Polajnar et al. 2014).

4. **Word prevalence as a new variable of word difficulty.** Thus far, word frequency has been used as the main proxy of word difficulty, for instance in algorithms to calculate text difficulty or to simplify texts. Figure 4 shows that this is only a crude approximation of word knowledge (Shardlow 2014). Based on the crowdsourcing data, we now have information about how many people know each word in the Dutch language. This will provide researchers with a better measure to estimate the level of difficulty of language samples (e.g., books and documents aimed at people with different proficiency levels). Similar attempts to collect word prevalence values for 'all' words are currently taking place for English and Spanish. The data will hopefully be released in the coming years.

5. Help from NLP researchers

At the same time, psycholinguists depend on input from NLP research to bring the megastudy approach to full fruition. Thus far, the input from NLP has been most prominent in the calculation of word frequencies. Although word form counting has its merits, the outcome is much richer and more interesting when it is accompanied by part-of-speech information (further described in Brysbaert et al. 2012). For some languages, such as Mandarin Chinese, this is even a bare necessity, given that the words are written without spaces (Cai and Brysbaert 2010). Part-of-speech information indicates which role each word plays in a sentence and requires the availability of high-quality automatic parsing, tagging and lemmatization algorithms. Another recent addition has been the calculation of N-gram counts (sequences of N words), which makes psycholinguistic research on the processing of multiword units possible (Arnon and Snider 2010, Baayen et al. 2011, Siyanova-Chanturia et al. 2011). Languages for which the required software is not (yet) available, are at a clear disadvantage in this respect.

Other NLP-based variables that are starting to have influence, are estimates of word and text similarity (e.g., Turney and Pantel 2010). For instance, Jones et al. (2012) argue that not so much word frequency matters for the speed of word recognition, but the semantic diversity of the contexts in which the word is encountered (see also Hoffman et al. 2013). This requires software to gauge the semantic similarity of texts. Other research looks at the semantic richness of words (Recchia and Jones 2012, Yap et al. 2011). Here too, automatically calculated measures of semantic richness and relatedness for large numbers of words are needed.

More in general, there is a high need in psycholinguistics for NLP measures of word meanings and word similarities, either based on the calculation of word co-occurrences or on initiatives such as Wordnet (Miller 1995). In the former approach, the meaning of words is gauged by analyzing the surrounding words. There are various techniques to do so (Mikolov et al. 2013, Turney and Pantel

2010) with increasing precision. One popular benchmark is to see how well the algorithms perform on a vocabulary test with multiple choice items. This test was introduced by Landauer and Dumais (Landauer and Dumais 1997) when they developed the Latent Semantic Analysis (LSA) model of word meanings. To see how well the algorithm worked, they tested for 80 words from a widely used multiple choice vocabulary test (TOEFL) how well the algorithm could predict the correct alternative among four choices. This was the case for 64.4% of the items, similar to human performance. Since, more powerful algorithms have been developed, which select the correct alternative for all items (Bullinaria and Levy 2012, see website). Wordnet is an electronic dictionary with entries of words organized in terms of their semantics. Specifically, words with related meanings are interlinked by means of pointers that stand for their semantic interrelationship. Wordnet was originally developed for English, but is becoming available for other languages as well (e.g., Black et al. 2006, Vossen et al. 1999).

A question that is being addressed with NLP measures of semantic similarity, is to what extent a limited number of word ratings (e.g., on AoA, concreteness, or affective values) can be used as seeds to automatically calculate the values of other words in the language. This is particularly interesting for subjective norms that are not yet available in large numbers, or for researchers working in languages without extensive norms. Bestgen and Vincze (2012) argued that on the basis of 1000 seed words, it is possible to calculate the valence of a new target word by averaging the valence of the 30 seed words that are semantically closest to the target word (determined on the basis of an LSA analysis). This work was recently extended by Mandera et al. (in press a) to other measures of semantic similarity and to other comparison methods. The authors also examined the usefulness of the estimates for psycholinguistic research.

It can be expected that the input from NLP to psycholinguistics will increase further, as more and more refined measures become available. In our experience, there is one serious limitation in the transfer of information, however. It is the tendency NLP researchers have not to make their measures available in an easy to read format. For most software engineers, the proof of concept is more important than the algorithm or the output of the algorithm. As a result, many potentially interesting measures never reach the psycholinguistic community and are never validated on human data. Certainly in the present age of massive information distribution, this is a missed opportunity. LSA-based semantic similarity (Landauer and Dumais 1997) is still the most frequently used NLP-type semantic variable in psycholinguistic research, not because the measure is the best available (see above), but because there is an easy-to-use website that calculates the measure for all words (and word combinations) in English. This makes the information not only available to technically literate users, but also to beginning psychology students. In the same vein, a text file with the output of an algorithm is much easier to use than the software itself (often provided without the corpus on which it operates). This is a small extra effort, which in our view would very much increase the impact of NLP on psycholinguistic research.

6. Conclusion

The turn to large datasets and megastudies in psycholinguistics provides new opportunities for collaboration with NLP researchers and computational linguists, because the skills and the data in the different domains are largely complementary. On the one hand, we hope to have shown that current technological and methodological developments make the collection of human data easier (and more affordable) than a few years ago. As a result, increasingly large datasets are created. These can serve as input or criterion variables for NLP research. At the same time, the output of NLP algorithms is interesting for psycholinguists. The chances of this information being used depends on whether it is made available in easy-to-use formats, which technically less proficient people can use.

References

Arnon, I. and N. Snider (2010), More than words: Frequency effects for multi-word phrases, *Journal of Memory and Language* **62**, pp. 67–82.

Baayen, R. H. (2010), A real experiment is a factorial experiment, *The Mental Lexicon* **5** (1), pp. 149–157.

Baayen, R. H., P. Milin, Hendrix P. Đurđević, D. F., and M. Marelli (2011), An amorphous model for morphological processing in visual comprehension based on naive discriminative learning, *Psychological Review* **118**, pp. 438–481.

Balota, D. A., M. J. Cortese, S. D. Sergent-Marshall, D. H. Spieler, and M. J. Yap (2004), Visual word recognition of single-syllable words, *Journal of Experimental Psychology: General* **133**, pp. 283–316.

Balota, D. A., M. J. Yap, K. A. Hutchison, and M. J. Cortese (2013), Megastudies: What do millions (or so) of trials tell us about lexical processing?, *in* Adelman, J. S., editor, *Visual Word Recognition Volume 1: Models and methods, orthography and phonology*, Psychology Press, New York, NY, pp. 90–115.

Balota, D. A., M. J. Yap, M. J. Cortese, K. A. Hutchison, B. Kessler, B. Loftis, J. H. Neely, D. L. Nelson, G. B. Simpson, and R. Treiman (2007), The English lexicon project, *Behavior Research Methods* **39**, pp. 445–459.

Bestgen, Y. and N. Vincze (2012), Checking and bootstrapping lexical norms by means of word similarity indexes, *Behavior Research Methods* **44** (4), pp. 998–1006.

Black, W., S. Elkateb, H. Rodriguez, M. Alkhalifa, P. Vossen, A. Pease, and C. Fellbaum (2006), Introducing the Arabic wordnet project, *Proceedings of the 3rd International WordNet Conference (GWC-06)*, pp. 295–299.

Brysbaert, M., A. B. Warriner, and V. Kuperman (2014a), Concreteness ratings for 40 thousand generally known English word lemmas, *Behavior Research Methods* **46**, pp. 904–911.

Brysbaert, M. and B. New (2009), Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English, *Behavior Research Methods* **41**, pp. 977–990.

Brysbaert, M., B. New, and E. Keuleers (2012), Adding Part-of-Speech information to the SUBTLEX-US word frequencies, *Behavior Research Methods* **44**, pp. 991–997.

Brysbaert, M., E. Keuleers, P. Mandera, and M. Stevens (2014b), *Woordenkennis van Nederlanders en Vlamingen anno 2013: Resultaten van het Groot Nationaal Onderzoek Taal*, Gent: Academia Press. (see also http://crr.ugent.be/archives/1494).

Brysbaert, M., M. Buchmeier, M. Conrad, A. M. Jacobs, J. Bölte, and A. Böhl (2011), The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German, *Experimental Psychology* **58**, pp. 412–424.

Brysbaert, M., M. Stevens, De Deyne, Voorspoels S., W., and G. Storms (2014c), Norms of age of acquisition and concreteness for 30, 000 Dutch words, *Acta Psychologica* **150**, pp. 80–84.

Bullinaria, J. A. and J. P. Levy (2012), Extracting semantic representations from word co-occurrence statistics: Stop-lists, stemming, and SVD, *Behavior Research Methods* **44**, pp. 890–907.

Cai, Q. and M. Brysbaert (2010), Subtlex-CH: Chinese word and character frequencies based on film subtitles, *PLoS ONE* **5** (6), pp. 1–8.

Cutler, A. (1981), Making up materials is a confounded nuisance, or: Will we able to run any psycholinguistic experiments at all in 1990?, *Cognition* **10** (1), pp. 65–70.

Dimitropoulou, M., J. A. Duñabeitia, A. Avilés, J. Corral, and M. Carreiras (2009), Subtitle-based word frequencies as the best estimate of reading behavior: the case of Greek, *Frontiers in psychology* **1**, pp. 218–218.

Ferrand, L., B. New, M. Brysbaert, E. Keuleers, P. Bonin, A. Meot, M. Augustinova, and C. Pallier (2010), The French Lexicon Project: Lexical decision data for 38, 840 French words and 38, 840 pseudowords, *Behavior Research Methods* **42**, pp. 488–496.

Gerhand, S. and C. Barry (1999), Age of acquisition, word frequency, and the role of phonology in the lexical decision task, *Memory and Cognition* **27** (4), pp. 592–602.

Gilhooly, K. J. and R. H. Logie (1980), Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1, 944 words, *Behavior Research Methods & Instrumentation* **12**, pp. 395–427.

Guerini, M., L. Gatti, and M. Turchi (2013), *Sentiment analysis: How to derive prior polarities from SentiWordNet*. http://arxiv.org/pdf/1309.5843.pdf.

Hill, F. and A. Korhonen (2014), Concretenss and subjectivity as dimensions of lexical meaning, *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, pp. 725–731. http://acl2014.org/acl2014/P14-2/pdf/P14-2118.pdf.

Hoffman, P., M. A. L. Ralph, and T. T. Rogers (2013), Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words, *Behavior Research Methods* **45**, pp. 718–730.

Hofland, K. and S. Johansson (1982), *Word frequencies in British and American English*, Bergen: Norwegian Computing Centre for the Humanities.

Jones, M. N., B. T. Johns, and G. Recchia (2012), The role of semantic diversity in lexical organization, *Canadian Journal of Experimental Psychology/Revue Canadienne de Psychologie Expérimentale* **66**, pp. 115–124.

Keuleers, E. and M. Brysbaert (2010), Wuggy: A multilingual pseudoword generator, *Behavior Research Methods* **42**, pp. 627–633.

Keuleers, E., K. Diependaele, and M. Brysbaert (2010a), Practice effects in large-scale visual word recognition studies: A lexical decision study on 14, 000 Dutch mono- and disyllabic words and nonwords, *Frontiers in Psychology* **1**, pp. 174.

Keuleers, E., M. Brysbaert, and B. New (2010b), Subtlex-NL: A new frequency measure for Dutch words based on film subtitles, *Behavior Research Methods* **42**, pp. 643–650.

Keuleers, E., M. Stevens, P. Mandera, and M. Brysbaert (in press), *Word knowledge in the crowd: Results of a massive online vocabulary test*.

Keuleers, E., P. Lacey, K. Rastle, and M. Brysbaert (2012), The British Lexicon Project: Lexical decision data for 28, 730 monosyllabic and disyllabic English words, *Behavior Research Methods* **44**, pp. 287–304.

Kuperman, V., H. Stadthagen-Gonzalez, and M. Brysbaert (2012), Age-of-acquisition ratings for 30, 000 English words, *Behavior Research Methods* **44** (4), pp. 978–990.

Kuperman, V., Z. Estes, M. Brysbaert, and A. B. Warriner (2014), Emotion and language: Valence and arousal affect word recognition, *Journal of Experimental Psychology: General* **143** (3), pp. 1065–1081.

Kučera, H. and W. N. Francis (1967), *Computational analysis of present-day American English*, Providence, RI: Brown University Press.

Landauer, T. K. and S. T. Dumais (1997), A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge, *Psychological Review* **104**, pp. 211–240.

Lemhöfer, K. and M. Broersma (2012), Introducing LexTALE: A quick and valid lexical test for advanced learners of English, *Behavior Research Methods* **44** (2), pp. 325–343.

Mandera, P., E. Keuleers, and M. Brysbaert (in press a), How useful are corpus-based methods for extrapolating psycholinguistic variables?, *Quarterly Journal of Experimental Psycholoogy*.

Mandera, P., E. Keuleers, Z. Wodniecka, and M. Brysbaert (in press b), Subtlex-pl: subtitle-based word frequency estimates for polish, *Behavior Research Methods*. http://link.springer.com/10.3758/s13428-014-0489-4.

Meara, P. and B. Buxton (1987), An alternative to multiple choice vocabulary tests, *Language Testing* **4** (2), pp. 142–154.

Mikolov, T., Q. V. Le, and I. Sutskever (2013), Exploiting Similarities among Languages for Machine Translation, *arXiv:1309*. http://arxiv.org/abs/1309.4168].

Miller, G. A. (1995), Wordnet: a lexical database for English, *Communications of the ACM* **38** (11), pp. 39–41.

Moors, A., De Houwer, Hermans J., Wanmaker D., van Schie S., Van Harmelen K., De Schryver A., De Winne M., J., and M. Brysbaert (2013), Norms of valence, arousal, dominance, and age of acquisition for 4300 Dutch Words, *Behavior Research Methods* **45**, pp. 169–177.

Morrison, C. M. and A. W. Ellis (1995), Roles of word frequency and age of acquisition in word naming and lexical decision, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **21** (1), pp. 116–133.

Muresan, I., A. Stan, M. Giurgiu, and R. Potolea (2013), Evaluation of sentiment polarity prediction using a dimensional and a categorical approach, *Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1–6. http://consortium.simple4all.org/files/2013/09/SentencePolarity_AS.pdf.

New, B., M. Brysbaert, J. Veronis, and C. Pallier (2007), The use of film subtitles to estimate word frequencies, *Applied Psycholinguistics* **28**, pp. 661–677.

Polajnar, T., L. Fagarasan, and S. Clark (2014), *Learning type-driven tensor-based meaning representations*. http://arxiv.org/pdf/1312.5985v2.pdf.

Recchia, G. and M. N. Jones (2012), The semantic richness of abstract concepts, *Frontiers in Human Neuroscience* **6**, pp. 1–16.

Ruppenhofer, J., M. Wiegand, and J. Brandes (2014), *Comparing methods for deriving intensity scores for adjectives*, EACL 2014. http://anthology.aclweb.org//E/E14/E14-4.pdf#page=137.

Shardlow, M. (2014), *Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline*, Reykjavik: LREC9. http://www.lrec-conf.org/proceedings/lrec2014/pdf/479_Paper.pdf.

Siyanova-Chanturia, A., K. Conklin, and W. J. van Heuven (2011), Seeing a phrase "time and again" matters: The role of phrasal frequency in the processing of multiword sequences, *Journal of Experimental Psychology: Learning, Memory, and Cognition* **37**, pp. 776–784.

Spieler, D. H. and D. A. Balota (1997), Bringing computational models of word naming down to the item level, *Psychological Science* **8**, pp. 411–416.

Sze, W. P., Rickard Liow, S. J., and M. J. Yap (2014), The Chinese Lexicon Project: a repository of lexical decision behavioral responses for 2, 500 Chinese characters, *Behavior Research Methods* **46** (1), pp. 263–73.

Turney, P. D. and P. Pantel (2010), From frequency to meaning: Vector space models of semantics, *Journal of Artificial Intelligence Research* **37**, pp. 141–188.

Vajjala, S. and D. Meurers (in press), *Readability Assessment for Text Simplification: From Analyzing Documents to Identifying Sentential Simplifications*.

van Heuven, W. J. B., P. Mandera, M. Keuleers, and M. Brysbaert (2014), Subtlex-UK: A new and improved word frequency database for British English, *Quarterly Journal of Experimental Psychology* **67**, pp. 1176–1190.

Vossen, P., L. Bloksma, and P. Boersma (1999), *The Dutch WordNet*, University of Amsterdam.

Warriner, A. B., V. Kuperman, and M. Brysbaert (2013), Norms of valence, arousal, and dominance for 13, 915 English lemmas, *Behavior Research Methods* **45** (4), pp. 1191–1207.

Yap, M. J., S. E. Tan, P. M. Pexman, and I. S. Hargreaves (2011), Is more always better? Effects of semantic richness on lexical decision, speeded pronunciation, and semantic classification, *Psychonomic Bulletin & Review* **18**, pp. 742–750.